

UNITED STATES PATENT & TRADEMARK OFFICE

Re: Application of: **Andreas SCHAALE**

Serial No.: To Be Assigned

Filed: Herewith as national phase of International
Application No. PCT/EP2004/003972, filed 15 April 2004

For: **METHOD FOR GENERATING DATA RECORDS
FROM A DATA BANK, ESPECIALLY FROM THE
WORLD WIDE WEB, CHARACTERISTIC SHORT
DATA RECORDS, METHOD FOR DETERMINING
DATA RECORDS FROM A DATA BANK WHICH
ARE RELEVANT FOR A PREDEFINED SEARCH
QUERY AND SEARCH SYSTEM FOR
IMPLEMENTING SAID METHOD**

Customer No: 23280

LETTER RE: PRIORITY

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

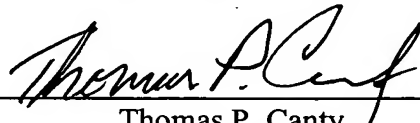
October 26, 2005

Sir:

Applicant hereby claims priority of German Application Serial No. 103 19 427.4, filed 29 April 2003, through International Application No. PCT/EP2004/003972, filed 15 April 2004.

Respectfully submitted,

DAVIDSON, DAVIDSON & KAPPEL, LLC

By 
Thomas P. Canty
Reg. No. 44,586

Davidson, Davidson & Kappel, LLC
485 Seventh Avenue, 14th Floor
New York, New York 10018
(212) 736-1940

BEST AVAILABLE COPY

EP04/03972

**PRIORITY
DOCUMENT**
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH RULE 17.1(a) OR (b)



REC'D 01 JUN 2004

WIPO

PCT

**Prioritätsbescheinigung über die Einreichung
einer Patentanmeldung**

Aktenzeichen:

103 19 427.4

Anmeldetag:

29. April 2003

Anmelder/Inhaber:

Contraco Consulting & Software Limited, London/GB

Bezeichnung:

Verfahren zur Erstellung von für Datensätze aus einer Datenbank, insbesondere aus dem World-Wide-Web, charakteristischen Kurzdatensätzen, Verfahren zur Ermittlung von für eine vorgebbare Suchanfrage relevanten Datensätzen aus einer Datenbank und Suchsystem zur Durchführung des Verfahrens

IPC:

G 06 F 17/30

Die angehefteten Stücke sind eine richtige und genaue Wiedergabe der ursprünglichen Unterlagen dieser Patentanmeldung.

München, den 22. April 2004
Deutsches Patent- und Markenamt

Der Präsident
Im Auftrag

Agurks

Beschreibung

Verfahren zur Erstellung von für Datensätze aus einer Datenbank, insbesondere aus dem World-Wide-Web, charakteristischen Kurzdatensätzen, Verfahren zur Ermittlung von für eine vorgebbare Suchanfrage relevanten Datensätzen aus einer Datenbank und Suchsystem zur Durchführung des Verfahrens

Die Erfindung bezieht sich auf ein Verfahren zur Erstellung von für Datensätze aus einer Datenbank, insbesondere aus dem World-Wide-Web, charakteristischen Kurzdatensätzen zur Hinterlegung auf einem Speichermodul als Basis zur Ermittlung der für eine vorgebbare Suchanfrage relevanten Datensätze. Sie betrifft weiter ein Verfahren zur Ermittlung von für eine vorgebbare Suchanfrage relevanten Datensätzen aus einer Datenbank, insbesondere aus dem World-Wide-Web, bei dem derartige Kurzdatensätze auf ihre Relevanz für die jeweilige Suchanfrage hin durchsucht werden. Weiterhin betrifft die Erfindung ein Suchsystem zur Ermittlung von für eine vorgebbare Suchanfrage relevanten Datensätzen aus einer Datenbank, insbesondere aus dem World-Wide-Web.

In komplexen Datenbanken oder auch im weltumspannenden Computernetz („World-Wide-Web“) wird eine enorme Menge an Informationen bereitgehalten, die ein Benutzer zu Recherchezwecken mehr oder weniger gezielt abrufen kann. Um dabei Informationen aus der großen Informationsmenge heraus gezielt nutzbar machen zu können, kommen so genannte Suchmaschinen zum Einsatz, die gerade bei der Informationsbeschaffung aus dem World-Wide-Web teilweise erhebliche Verbreitung erfahren haben. Bei derartigen Suchmaschinen wird dem Benutzer üblicherweise über ein Ein-/Ausgabemodul ein Anfragenfenster zur Verfügung gestellt, über das gezielte Such- oder Recherchebegriffe vorgegeben werden können. Anschließend durchsucht die Suchmaschine die Informationsbasis der Datenbank bzw. des World-Wide-Web nach geeigneten Schlüssel- oder Keywörtern. Die daraufhin aufgefundenen Antwortdatensätze werden üblicherweise von der jeweiligen Suchmaschine hinsichtlich ihrer Relevanz für den vorgegebe-

1 nen Suchauftrag kategorisiert und dem Benutzer in der Art einer Trefferliste in
2 nach ihrer Relevanz geordneter Reihenfolge zur Verfügung gestellt.

3
4 Allerdings bereitet die zunehmende Komplexität der eingesetzten Datenbanken
5 und insbesondere der stetig wachsende enorme Informationsgehalt im World-
6 Wide-Web zunehmende Probleme bei der strukturierten und gezielten Beschaf-
7 fung von Informationen. Die für die Recherche eingesetzten Suchmaschinen wer-
8 den daher zunehmend bezüglich der eingesetzten Suchalgorithmen verbessert,
9 wobei zudem in der Art von Vorsortierungen oder Vorfilterungen weitere Hilfsmittel
10 zur Klassifizierung von Datensätzen aus der Datenbank zum Einsatz kommen
11 können. Insbesondere im World-Wide-Web sind die Datensätze üblicherweise in
12 Form von so genannten Domains strukturiert und organisiert, wobei eine Domain
13 typischerweise von einem Betreiber unterhalten wird und ihrerseits eine Vielzahl
14 von Unterdatensätzen, Textdokumenten oder dergleichen umfassen kann.

15
16 Um gerade bei der Informationsbeschaffung aus dem World-Wide-Web trotz der
17 enormen Vielzahl der dort verfügbaren Datensätze oder Domains mit einem sei-
18 nerseits möglicherweise vergleichsweise großen Inhalt an Unterdatensätzen oder
19 Informationsträgern eine geeignete Vorauswahl von bei einer vorgegebenen
20 Suchanfrage zu berücksichtigenden Domains treffen zu können, kann ein so ge-
21 nanntes Ranking der Domains zum Einsatz kommen. Dabei wird jeder Domain ein
22 Kennwert zugeordnet, der anhand von zugänglichen Sekundärinformationen in
23 der Art einer relativen Relevanz die Wichtigkeit der Berücksichtigung der jeweili-
24 gen Domain für die Suchanfrage charakterisiert. Bei der Zuordnung dieses Kenn-
25 werts zur jeweiligen Domain wird üblicherweise in der Art eines so genannten sta-
26 tischen Ansatzes eine Informationsbasis herangezogen, bei der beispielsweise
27 aufgrund des Grades der Vernetzung der jeweiligen Domain mit anderen Domains
28 auf die relative Bedeutung der jeweiligen Domain geschlossen wird. Als Maß für
29 eine derartige Bedeutung kann die Anzahl der so genannten Links oder Querver-
30 weise von anderen Domains auf die jeweilige Domain herangezogen werden, wo-
31 bei die Annahme zugrunde gelegt wird, dass eine Vielzahl von Querverweisen auf
32 die jeweilige Domain ein Hinweis darauf ist, dass diese Domain für eine Vielzahl

1 von Benutzern von besonderer Wichtigkeit bei der Bearbeitung ihrer Suchanfra-
2 gen ist.

3
4 Allerdings hat sich gezeigt, dass bei der Zuordnung einer derartigen statischen
5 Kenngröße als Relevanzindikator für die jeweilige Domain Spielraum für Manipula-
6 tionen besteht, wobei unabhängig von den eigentlichen Benutzerinteressen auf-
7 grund wirtschaftlicher Entscheidungen durch die Herstellung einer Vielzahl sach-
8 lich eigentlich nicht gerechtfertigter Links oder Querverweise künstlich der Ein-
9 druck einer vergleichsweise hohen Relevanz oder Bedeutsamkeit der jeweiligen
10 Domain erzeugt werden kann. Die Verwendung derartiger statischer Relevanzzu-
11 ordnungen zur Verbesserung der Suchergebnisse von Internet-Recherchen ist da-
12 her von eher abnehmendem Interesse.

13
14 Die enormen, im World-Wide-Web oder Internet verfügbaren Informationsmengen
15 machen bei der Durchführung einer Informationsrecherche die tatsächliche Durch-
16 suchung sämtlicher Domains inklusive der darin enthaltenen Unterdatensätze
17 oder Textbausteine in Echtzeit auf das Vorhandensein der Suchanfrage oder von
18 einzelnen Elementen der jeweiligen Suchanfrage unmöglich. Statt dessen kom-
19 men in Suchsystemen oder Suchmaschinen für die Informationsbeschaffung aus
20 dem Internet oder World-Wide-Web so genannte „Crawler“ oder Durchsuchermo-
21 dule zum Einsatz, die ständig die Domains oder Datensätze aus dem World-Wide-
22 Web bzw. einer komplexen Datenbank auf ihren Textinhalt oder andere als rele-
23 vant erachtete Informationen hin durchsuchen. Im Rahmen von vorgegebenen
24 Systemressourcen (beispielsweise Bearbeitungszeit, Speicherkapazität oder Re-
25 chenleistung) durchsucht das jeweilige Durchsuchermodul die gerade angewählte
26 Domain oder den Datensatz und stellt bis zu einer durch die zugewiesenen Sys-
27 temressourcen vorgegebenen Grenze anhand der in der jeweiligen Domain aufge-
28 fundenen Informationen einen für die Domain oder den Datensatz charakteristi-
29 schen Kurzdatensatz, beispielsweise in Form einer Textdatei mit möglicherweise
30 zugeordneten Überschriften oder sonstigen Indikatoren, zusammen.

Dieser Kurzdatensatz wird sodann auf einem Speichermodul hinterlegt und für eine nachfolgende Untersuchung bereitgehalten. Die Gesamtheit der aus den überhaupt bei diesem Vorgehen berücksichtigten Datensätze oder Domains erstellten, auf dem Speichermodul hinterlegten Kurzdatensätze wird auch als so genannter „Index“ der jeweiligen Suchmaschine bezeichnet und dient als Informationsbasis für die nachfolgend durchgeführten Recherchen. Die Erstellung der im Index enthaltenen Kurzdatensätze erfolgt dabei üblicherweise kontinuierlich, wobei individuelle Domains zyklisch angewählt werden, so dass eine andauernde Aktualisierung des Index erfolgt. Bei einer nachfolgenden Recherche, also bei der Ermittlung der für eine vorgegebene Suchanfrage relevanten Datensätze, wird sodann der durch die Gesamtheit der hinterlegten Kurzdatensätze gebildete Index auf das Vorhandensein von Schlüssel- oder Keywörtern der jeweiligen Suchanfrage oder von Elementen davon durchsucht, wobei anhand der dabei erzielten Suchergebnisse oder Treffer die den aufgefundenen Kurzdatensätzen zugeordneten Datensätze oder Domains als relevant für die jeweilige Suchanfrage ermittelt werden.

Aufgrund der Vielzahl der im World-Wide-Web verfügbaren Domains oder Datensätze können nicht sämtliche Domains bei der Erstellung der Kurzdatensätze berücksichtigt werden. Die Entscheidung, welche Domains bei der Erstellung des Index berücksichtigt werden, erfolgt dabei üblicherweise anhand der oben genannten Relevanzkriterien, also insbesondere anhand von Informationen über eine erkannte oder vermeintliche Bedeutung der jeweiligen Domain für die Benutzer. Gerade da infolge der verfügbaren Informationsvielfalt eine besonders zielgerichtete Vorsortierung der Informationen und insbesondere der als für eine Suchanfrage relevant erkannten Datensätze für die nachfolgende Auswertung der Rechercheergebnisse sehr bedeutsam sein kann, ist bereits bei der Erstellung des so genannten Index eine besonders hohe Qualität und Sorgfalt bei der Auswertung der berücksichtigten Informationen wünschenswert.

Der Erfindung liegt daher die Aufgabe zugrunde, ein Verfahren zur Erstellung von für Datensätze charakteristischen Kurzdatensätzen der oben genannten Art anzugeben, mit dem ein für eine qualitativ hochwertige Informationsbeschaffung aus

1 der Datenbank oder aus dem World-Wide-Web besonders geeigneter Suchindex
2 erzeugt werden kann. Des Weiteren sollen unter Verwendung dieses Verfahrens
3 ein besonders geeignetes Verfahren zur Ermittlung von für eine vorgebbare Such-
4 anfrage relevanten Datensätzen aus einer Datenbank, insbesondere aus dem
5 World-Wide-Web, und ein Suchsystem zur Durchführung dieses Verfahrens ange-
6 geben werden.

7
8 Bezüglich des Verfahrens zur Erstellung der für die Datensätze charakteristischen
9 Kurzdatensätze wird diese Aufgabe erfindungsgemäß gelöst, indem die zur Erstel-
10 lung eines Kurzdatensatzes aus einem Datensatz bereitgestellten Systemressour-
11 cen unter Berücksichtigung von ermittelten Erfahrungswerten aus vorangegange-
nen Suchanfragen gewählt werden.

13
14 Die Erfindung geht dabei von der Überlegung aus, dass für die Erzeugung einer
15 für qualitativ besonders hochwertige Informationsbeschaffung besonders geeig-
16 neten Informationsbasis auf der Grundlage der für die Datensätze charakteristi-
17 schen Kurzdatensätze zwar einerseits in der Art statischer Kennwerte verfügbare
18 Informationen über die einzelnen Datensätze oder Domains an sich berücksichtigt
19 werden können, das aber andererseits in der Art eines dynamischen Elements
20 auch für die Benutzerinteressen charakteristische Informationen mitberücksichtigt
21 werden sollten. Dabei liegt die Erkenntnis zugrunde, dass das Ergebnis einer In-
22 formationsbeschaffung aus der Datenbank oder dem World-Wide-Web dann als
qualitativ besonders hochwertig angesehen wird, wenn es in möglichst weit rei-
24 chendem Umfang das Benutzerinteresse korrekt widerspiegelt. Es sollten daher
25 Maßnahmen getroffen werden, um für das Benutzerinteresse charakteristische In-
26 formationen in die weitere Informationsbeschaffung einfließen zu lassen. Ein An-
27 satzpunkt hierfür ist bereits die Erzeugung der Informationsbasis für die Bearbei-
28 tung der Suchanfragen, so dass sich Informationen über das Besucherinteresse
29 bereits bei der Erzeugung der für die Datensätze oder Domains charakteristischen
30 Kurzdatensätze im Index niederschlagen sollten. Um dies zu ermöglichen, werden
31 bereits bei der Zuweisung der Systemressourcen, die für die Erstellung eines
32 Kurzdatensatzes aus einem zugeordneten Datensatz aufgewendet werden kön-

1 nen und somit die Vollständigkeit der im Kurzdatensatz für die Informationsbe-
2 schaffung bereitgehaltenen Informationen entscheidend mitbestimmen, im Hin-
3 blick auf das Benutzerinteresse unter Berücksichtigung von Erfahrungswerten aus
4 vorangegangenen Suchanfragen zugewiesen.

5
6 In besonders einfacher und prägnanter Weise können die Benutzerinteressen be-
7 reits bei der Erstellung des Index berücksichtigt werden, indem vorteilhafterweise
8 bei der Zuordnung der Systemressourcen die Häufigkeit von zu einer Suchanfrage
9 gleichen oder ähnlichen Suchanfragen in der jüngsten Vergangenheit als Erfah-
10 rungswert berücksichtigt wird. In weiterer vorteilhafter Ausgestaltung kann dabei
11 auch die Trefferhäufigkeit der Datensätze oder Domains bzgl. der von den Benut-
12 zern in jüngster Vergangenheit besonders häufig vorgegebenen Suchanfragen be-
13 rücksichtigt werden. Die Erfahrungswerte umfassen daher zweckmäßigerweise ei-
14 ne für die Anzahl gleichartiger Suchanfragen in einem vorgebbaren Zeitintervall
15 charakteristische Kennzahl.

16
17 Um die Benutzerinteressen besonders zielführend bereits bei der Erstellung des
18 Index für die Suchmaschine berücksichtigen zu können, werden vorteilhafterweise
19 als Systemressourcen die Ressourcen eines zur Erstellung der für die Datensätze
20 jeweils charakteristischen Kurzdatensätze vorgesehenen Durchsuchermoduls
21 oder Crawlers unter Berücksichtigung von ermittelten Erfahrungswerten aus vor-
22 angegangenen Suchanfragen gewählt.

23
24 In einer besonders vorteilhaften Ausgestaltung werden die Benutzerinteressen bei
25 der Zuordnung der Systemressourcen besonders weitgehend berücksichtigt, in-
26 dem bei der Ermittlung der Erfahrungswerte in besonderem Maß die möglicher-
27 weise komplexe Struktur der von den Benutzern verwendeten Suchanfragen be-
28 rücksichtigt wird. Dabei liegt die Erkenntnis zugrunde, dass ein besonders genau-
29 es Abbild des allgemeinen Benutzerinteresses nicht allein durch die relative Häu-
30 figkeit einzelner in Suchanfragen verwendeter Elemente oder Begriffe, sondern er-
31 gänzend oder zusätzlich auch noch durch die Berücksichtigung von spezifischen
32 Korrelationen zwischen einzelnen Begriffen oder Elementen von Suchanfragen er-

1 zielt werden kann. Dabei wird insbesondere berücksichtigt, dass einzelne Elemen-
2 te oder Komponenten einer Suchanfrage entsprechend dem aktuell gerade ver-
3 breiteten Benutzerinteresse bevorzugt in Kombination mit spezifischen anderen
4 Einzelelementen oder Komponenten von Suchanfragen angefragt werden. Bei-
5 spielsweise könnte das aktuelle Benutzerinteresse im Allgemeinen aktuell in eine
6 Richtung gehen, dass bevorzugt kostenlose Multimedia-Dateien aus dem Internet
7 heruntergeladen werden sollen. In einer derartigen Konstellation ist bei Suchanfra-
8 gen vermehrt mit einer Kombination der Suchbegriffe „MP3“, „free“ und „down-
9 load“ zu rechnen. Bei der spezifischen Bewertung und Berücksichtigung vergan-
10 gener Suchanfragen kann somit die Kombination dieser drei Einzelelemente einer
11 Suchanfrage als ein besonders gewichtiger Indikator für gesteigertes Benutzerin-
12 teresse gewertet werden. Um dies zu ermöglichen, werden bei der Ermittlung der
13 Erfahrungswerte vorzugsweise Korrelationen zwischen Einzelelementen der Such-
14 anfragen berücksichtigt.

15
16 Um in der Art einer ersten Vorfilterung eine vergleichsweise einfach beschaffbare
17 Eingangsinformation für die Bewertung von Suchanfragen und ihrer Relevanz für
18 die Datensätze bereitzustellen, wird bei der Ermittlung der Erfahrungswerte vorteil-
19 hafterweise die relative Häufigkeit von Suchanfragen und/oder von Einzelelemen-
20 ten der Suchanfragen berücksichtigt. Dies kann auf besonders einfache Weise
21 unmittelbar bei der Erstellung des Index berücksichtigt werden, indem denjenigen
22 Datensätzen, die für eine vorgegebene Suchanfrage oder für eine vorgegebene
23 Kombination aus Einzelelementen von Suchanfragen als relevant erkannt werden,
24 vorteilhafterweise in Abhängigkeit von der relativen Häufigkeit der Suchanfrage
25 bzw. der Kombination aus Einzelelementen von Suchanfragen zusätzliche Sys-
26 temressourcen für die Erstellung des zugeordneten Kurzdatensatzes zugewiesen
27 werden.

28
29 Vorteilhafterweise werden die auf die genannte Art erstellten, für die Datensätze
30 aus der Datenbank charakteristischen Kurzdatensätze zur Ermittlung von für eine
31 vorgebbare Suchanfrage relevanten Datensätzen aus der Datenbank, insbeson-
32 dere aus dem World-Wide-Web, genutzt, indem die solchermaßen erstellten, in

1 einem Speichermodul hinterlegten Kurzdatensätze auf ihre Relevanz für die jewei-
2 lige Suchanfrage hin durchsucht werden. Als Kriterium für die Ermittlung dieser
3 Relevanz kann dabei beispielsweise die Häufigkeit dienen, mit der ein Schlüssel-
4 oder Keyword der Suchanfrage im jeweiligen Kurzdatensatz aufgefunden werden
5 kann, wobei zudem noch eine Differenzierung nach dem Ort des jeweiligen Auffin-
6 dens, beispielsweise in einer Überschrift oder im Volltext, vorgenommen werden
7 kann.

8
9 Bezüglich des Suchsystems zur Ermittlung von für eine vorgebbare Suchanfrage
10 relevanten Datensätzen aus einer Datenbank, insbesondere aus dem World-
11 Wide-Web, wird die genannte Aufgabe gelöst, indem in einem Speichermodul für
12 die Datensätze charakteristische Kurzdatensätze hinterlegt sind, wobei die zur Er-
13 stellung eines Kurzdatensatzes aus einem Datensatz bereitgestellten Systemres-
14 sourcen unter Berücksichtigung von hinterlegten Erfahrungswerten aus vorange-
15 gangenen Suchanfragen gewählt sind.

16
17 Die Erfahrungswerte umfassen dabei vorteilhafterweise eine für die Anzahl gleich-
18 artiger Suchanfragen in einem vorgebbaren Zeitintervall charakteristische Kenn-
19 zahl. In zusätzlicher oder alternativer vorteilhafter Ausgestaltung sind als System-
20 ressourcen die Ressourcen eines zur Erstellung der für die Datensätze jeweils
21 charakteristischen Kurzdatensätze vorgesehenen Durchsuchermoduls unter Be-
22 rücksichtigung von hinterlegten Erfahrungswerten aus vorangegangenen Suchan-
fragen gewählt.

23
24
25 Die mit der Erfindung erzielten Vorteile bestehen insbesondere darin, dass durch
26 die Berücksichtigung von Erfahrungswerten aus vorangegangenen Suchanfragen
27 bei der Vergabe der Systemressourcen bei der Erstellung des Index oder der für
28 die Datensätze charakteristischen Kurzdatensätze bereits in einem besonders frü-
29 hen Stadium, nämlich in der Vorbereitungsphase einer Datenbank- oder Internet-
30 Recherche, eine weit gehende Berücksichtigung des aktuell vorhandenen Benut-
31 zerinteresses ermöglicht ist. Gerade durch die Berücksichtigung des Benutzerin-
32 teresses ergänzend zu oder anstelle von bisher verwendeten Datenbank spezifi-

1 schen Charakteristika wie beispielsweise der Häufigkeit der jeweiligen Querver-
2 weise ist eine vom Benutzer als besonders hochwertig angesehene Informations-
3 beschaffung ermöglicht. Ein besonders spezifisches Abbild des Benutzerinteres-
4 ses und somit eine besonders hohe Genauigkeit bei der Vergabe der Ressourcen
5 ist dabei erreichbar durch die Berücksichtigung von Korrelationen zwischen einzel-
6 nen Elementen von Suchanfragen, wobei gerade besonders häufig verwendete
7 Kombinationen spezifischer Einzelelemente und der Rückschluss auf die mit der-
8 artigen kombinierten Suchanfragen als Ergebnis aufgefundenen Datensätze oder
9 Domains eine in besonderem Maße an das Benutzerinteresse angepasste Tref-
10 fergenerierung erwarten lässt.

11
12 Ein Ausführungsbeispiel der Erfindung wird anhand einer Zeichnung näher erläu-
13 tert. Darin zeigt die Figur schematisch ein Suchsystem zur Ermittlung von für eine
14 vorgebbare Suchanfrage relevanten Datensätzen oder Domains aus dem World-
15 Wide-Web.

16
17 Das Suchsystem 1 gemäß der Figur ist über die mittels der Doppelpfeile 2 ange-
18 deuteten Datenleitungen des Internets oder World-Wide-Web mit einer Vielzahl
19 von Domains 4 verbunden, wobei jede Domain 4 ihrerseits typischerweise eine
20 Vielzahl von Unterdatensätzen, Textbausteinen, multimedialen Informationsele-
21 menten oder dergleichen umfasst.

22
23 Aufgrund der großen Vielzahl an im World-Wide-Web verfügbaren Informationen
24 ist das Suchsystem 1 für die Bearbeitung einer Suchanfrage nicht für die Durch-
25 suchung der Domains 4 oder der in ihnen enthaltenen Informationsinhalte auf das
26 Vorhandensein bestimmter Schlüsselwörter oder Keywords, sondern statt dessen
27 für die Durchsuchung eines in einem Speichermodul 6 hinterlegten so genannten
28 Index 8 ausgelegt. Der Index 8 umfasst eine Vielzahl von Kurzdatensätzen 10,
29 von denen jeder jeweils für einen Datensatz oder eine Domain 4 des World-Wide-
30 Web charakteristisch ist. Jeder Kurzdatensatz 10 enthält dabei einen als relevant
31 erkannten Teil des Informationsgehalts der jeweils zugeordneten Domain 4, wobei
32 im Kurzdatensatz 10 insbesondere die in der jeweiligen Domain 4 enthaltene

1 Textinformation wiedergegeben ist. Zur Bearbeitung einer Suchanfrage wird die-
2 se, wie durch den Pfeil 12 angedeutet, einem Ein-/Ausgabemodul 14 des Suchsys-
3 tems 1 zugeführt, von wo aus anhand von für die Suchanfrage charakteristischen
4 Schlüssel- oder Keywörtern eine Durchsuchung der Kurzdatensätze 10 gestartet
5 wird. Abhängig von der Anzahl der Ergebnisse oder Treffer, mit denen das Vor-
6 handensein von Schlüssel- oder Keywörtern in den Kurzdatensätzen 10 festge-
7 stellt wird, wird die mit dem jeweiligen Kurzdatensatz 10 korrespondierende Do-
8 main 4 als für die Suchanfrage relevant erkannt und die entsprechende Domain-
9 adresse dem Benutzer auf einer Ergebnisliste mitgeteilt.

10
11 Zur Erstellung der für die Domains 4 charakteristischen Kurzdatensätze 10, die in
12 ihrer Gesamtheit den Index 8 bilden, umfasst das Suchsystem 1 ein auch als
13 „Crawler“ bezeichnetes Durchsuchermodul 16. Das Durchsuchermodul 16 nimmt
14 in regelmäßigen, vorzugsweise zyklischen Zeitintervallen Kontakt mit den jeweili-
15 gen Domains 4 auf und durchsucht diese hinsichtlich ihres Informationsgehalts.
16 Dabei kann insbesondere vorgesehen sein, die auf der jeweiligen Domain 4 hin-
17 terlegte Textinformation zu erfassen und geeignet zu komprimieren. Art und Um-
18 fang der Analyse des Inhalts jeder Domain 4 durch das Durchsuchermodul 16
19 werden dabei durch die Vorgabe spezifischer Systemressourcen des Durchsu-
20 chermoduls 16 für die jeweilige Domain 4 festgelegt. Als Systemressourcen kön-
21 nen dabei abhängig von der jeweiligen Domain 4 beispielsweise die für die Durch-
22 suchung vorgesehene Zeitspanne, die eingesetzte Rechnerleistung und/oder zu-
23 gewiesene Speicherkapazitäten vorgegeben sein. Dabei kann insbesondere auch
24 vorgegeben sein, ob die jeweilige Domain 4 vom Durchsuchermodul 16 überhaupt
25 angesprochen oder von vornherein ignoriert werden soll. Anhand der bei der
26 Durchsuchung ermittelten Informationsbasis für die jeweilige Domain 4 wird an-
27 schließend vom Durchsuchermodul 16 in der Art einer Kurzfassung der zugeord-
28 nete Kurzdatensatz 10 erstellt und als Bestandteil des Index 8 im Speichermodul
29 6 abgelegt.

30
31 Die Zuordnung der Systemressourcen für die Durchsuchung der jeweiligen Do-
32 main 4 kann beispielsweise in Abhängigkeit domainspezifischer Relevanzkenn-

1 werte erfolgen. Dabei können auch so genannte statische Relevanzkennwerte
2 vorgesehen sein, die anhand vorgegebener Kriterien wie beispielsweise dem Ver-
3 netzungsgrad einer Domain 4 mit anderen Domains 4 charakterisieren, wie hoch
4 der Akzeptanzgrad der jeweiligen Domain 4 bei den Benutzern ist. Anhand dieser
5 Relevanzkennwerte kann festgelegt sein, ob eine Domain 4 bei der Durchsuchung
6 überhaupt berücksichtigt wird, und falls ja, wie sorgfältig die Durchsuchung der je-
7 weiligen Domain 4 bei der Erstellung des zugeordneten Kurzdatensatzes 10 erfol-
8 gen soll.

9
10 Darüber hinaus ist das Suchsystem 1 aber auch dafür ausgelegt, bei der Erstel-
11 lung der Kurzdatensätze 10 Erfahrungswerte und Erkenntnisse aus den vorange-
12 gangenen Suchanfragen mit zu berücksichtigen und somit das darin wiedergege-
13 bene aktuelle Benutzerinteresse in besonderem Maße in die Erstellung oder zykli-
14 sche Erneuerung des Index 8 miteinfließen zu lassen. Dazu ist dem Speichermo-
15 dul 6 ein weiteres Speichermodul 18 zugeordnet, in dem in der Art eines Log-
16 buchs die eingehenden Suchanfragen zur weiteren Auswertung abgelegt werden.
17 Die Inhalte des Speichermoduls 18 werden einem Analysemodul 20 zugänglich
18 gemacht, das die eingegangenen Suchanfragen einer Auswertung unterzieht und
19 anhand der dabei gewonnenen Erkenntnisse eine Neuverteilung der Systemres-
20 sourcen auf die zu berücksichtigenden Domains 4 beim nächsten Durchsuchungs-
21 zyklus vornimmt. Die entsprechende Zuordnung der Systemressourcen übermittelt
22 das Analysemodul 20, wie über den Pfeil 22 dargestellt, anschließend an das
Durchsuchermodul 16.

24
25 Bei der Zuordnung der Systemressourcen berücksichtigt das Analysemodul 20 so-
26 mit Erfahrungswerte aus vorangegangenen Suchanfragen. Dies kann beispiels-
27 weise dadurch geschehen, dass die Häufigkeit einer Suchanfrage oder eines
28 Schlüssel- oder Keywords als Einzelelement einer Suchanfrage ermittelt wird, wo-
29 bei bei häufig verwendeten Suchanfragen oder Einzelelementen von Suchanfra-
30 gen auf eine derzeit vergleichsweise hohe Popularität bei den Benutzern ge-
31 schlossen wird. Dementsprechend wird angenommen, dass die bei vergleichswei-
32 se populären Suchanfragen aufgefundenen und als relevant erkannten Datensät-

1 ze oder Domains 4 in vergleichsweise hohem Maße das aktuelle Benutzerinteres-
2 se widerspiegeln. In dieser Ausführungsform kann das Analysemodul 20 somit
3 denjenigen Domains 4, die als Ergebnis für vergleichsweise häufig verwendete
4 Suchanfragen aufgeführt werden, einen entsprechend erhöhten Anteil an System-
5 ressourcen bei der nächstfolgenden Durchsuchung durch das Durchsuchermodul
6 16 zuordnen.

7
8 Darüber hinaus ist das Suchsystem 1 aber auch dafür ausgelegt, vergleichsweise
9 komplexe Strukturen im Profil der Suchanfragen bei der Zuordnung der System-
10 ressourcen durch das Analysemodul 20 mitzuberücksichtigen. Dabei werden bei
11 der Ermittlung der Erfahrungswerte insbesondere auch Korrelationen zwischen
12 Einzelementen von Suchanfragen mitberücksichtigt. Falls beispielsweise festge-
13 stellt wird, dass einzelne Elemente oder Suchwörter in Suchanfragen besonders
14 häufig mit bestimmten anderen Einzelementen oder Suchwörtern kombiniert
15 werden, so wird auf eine hohe intrinsische Korrelation zwischen diesen beiden
16 Suchelementen geschlossen, so dass einerseits diejenigen Domains 4, in denen
17 vollständige oder approximative Kombinationen aufgefunden werden, als
18 besonders relevant erkannt werden, wobei andererseits bei der Auswertung der
19 relativen Häufigkeiten einzelner Suchelemente auch die relativen Häufigkeiten der
20 damit besonders korrelierten weiteren Suchelemente berücksichtigt werden
21 können.

22
23 Zur statistischen Auswertung der Suchanfragen wird dabei im Analysemodul 20
24 eine Korrelationsmatrix erstellt, deren Matrixelemente ein quantitatives Maß für die
25 Korrelation zwischen jeweils zwei Einzelementen von Suchanfragen angeben.
26 Als quantitatives Maß kann dabei insbesondere die relative Häufigkeit, mit der die
27 beiden jeweiligen Einzelemente von Suchanfragen miteinander in Kombination
28 gefragt werden, vorgesehen sein. Diese Korrelationsmatrix wird anschließend
29 durch eine Hauptachsentransformation diagonalisiert, wobei auf der Hauptdiago-
30 nalen der diagonalisierten Matrix die Eigenwerte der ursprünglichen Korrelations-
31 matrix angegeben sind. Bei dieser Hauptachsentransformation werden zudem die
32 Eigenvektoren der Korrelationsmatrix ermittelt.

Die Eigenwerte und Eigenvektoren der Korrelationsmatrix können anschließend für eine weitere Auswertung der Suchanfragen herangezogen werden. Diejenigen Eigenvektoren der Korrelationsmatrix, die einen vergleichsweise großen Eigenwert aufweisen, entsprechen dabei einem Mix von Einzelelementen von Suchanfragen, der gemäß der Linearkoeffizienten der Einzelelemente der Suchanfragen vergleichsweise häufig in typischen Suchanfragen vorkommt und somit in besonderem Maße das aktuelle Benutzerinteresse widerspiegelt. Es werden somit in einem nächsten Schritt diejenigen Eigenvektoren der Korrelationsmatrix ausgewählt, denen ein vergleichsweise großer Eigenwert zugeordnet ist. Durch die so ermittelten Eigenvektoren erhält man im Ergebnis einen Mix von Suchanfragen, die in der jeweiligen Kombination mit besonders hoher Wahrscheinlichkeit in jüngster Vergangenheit vorgekommen sind.

Mit den so ausgewählten, vergleichsweise großen Eigenwerten der Korrelationsmatrix zugeordneten „Eigenanfragen“ greift das Analysemodul 20 in der Art einer Testanfrage auf den Index 8 zu und ermittelt so für jede „Eigenanfrage“ die zu dieser Eigenanfrage als relevant erkannten Datensätze oder Domains 4. Da die solchermaßen ermittelten Domains 4 in besonderem Maße dem aktuellen Benutzerinteresse entsprechen, werden für diese Domains 4 die Systemressourcen bei der erneuten Durchsuchung des World-Wide-Web im Vergleich zum vorigen Durchlauf proportional erhöht. Dies kann beispielsweise durch eine Zuweisung eines Wichtungsfaktors bei der Bereitstellung der Systemressourcen für die jeweilige Domain 4 nach der Beziehung

$$R_{VPA}(D_k) = (1 + \alpha * \lambda_k^\beta) \alpha, \beta > 0$$

erfolgen, wobei λ_k der Eigenwert der zugehörigen Eigenanfrage D_k eine auf diese Eigenanfrage als Treffer angezeigte Domain 4 und α eine geeignet gewählte Konstante >0 sein kann.

Zusammenfassung

Ein Verfahren zur Erstellung von für Datensätze (4) aus einer Datenbank, insbesondere aus dem World-Wide-Web, charakteristischen Kurzdatensätzen (10) zur Hinterlegung auf einem Speichermodul (6) als Basis zur Ermittlung der für eine vorgebbare Suchanfrage relevanten Datensätze (4) soll zur Erzeugung eines für eine qualitativ hochwertige Informationsbeschaffung aus der Datenbank oder aus dem World-Wide-Web besonders geeigneter Suchindex geeignet sein. Dazu werden erfindungsgemäß die zur Erstellung eines Kurzdatensatzes (10) aus einem Datensatz (4) bereitgestellten Systemressourcen unter Berücksichtigung von ermittelten Erfahrungswerten aus vorangegangenen Suchanfragen gewählt.

Figur

Bezugszeichenliste

- 1 Suchsystem
- 2 Doppelpfeile
- 4 Domain
- 6 Speichermodul
- 8 Index
- 10 Kurzdatensätze
- 12 Pfeil
- 14 Ein-/Ausgabemodul
- 16 Durchsuchermodul
- 18 Speichermodul
- 20 Analysemodul
- 22 Pfeil

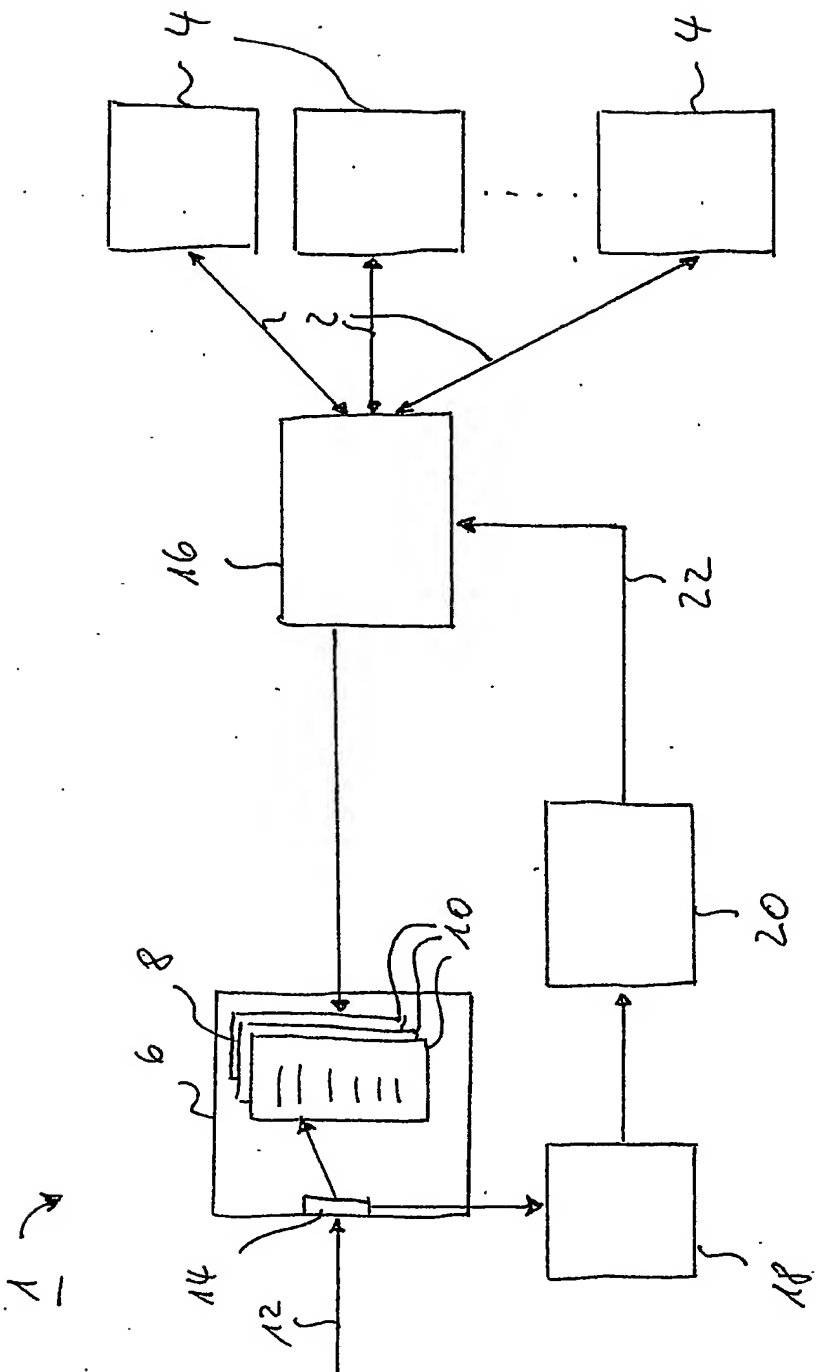
1 P030480DE-6/20

2
3 29. April 2003

4
5
6 Ansprüche

- 7
- 8 1. Verfahren zur Erstellung von für Datensätze (4) aus einer Datenbank, insbe-
9 sondere aus dem World-Wide-Web, charakteristischen Kurzdatensätzen
10 (10) zur Hinterlegung auf einem Speichermodul (6) als Basis zur Ermittlung
11 der für eine vorgebbare Suchanfrage relevanten Datensätze (4), bei dem die
12 zur Erstellung eines Kurzdatensatzes (10) aus einem Datensatz (4) bereitge-
13 stellten Systemressourcen unter Berücksichtigung von ermittelten Erfah-
14 rungswerten aus vorangegangenen Suchanfragen gewählt werden.
15
 - 16 2. Verfahren nach Anspruch 1, bei dem die Erfahrungswerte eine für die Anzahl
17 gleichartiger Suchanfragen in einem vorgebbaren Zeitintervall charakteristi-
18 sche Kennzahl umfassen.
19
 - 20 3. Verfahren nach Anspruch 1 oder 2, bei dem als Systemressourcen die Res-
21 sourcen eines zur Erstellung der für die Datensätze (4) jeweils charakteristi-
22 schen Kurzdatensätze (10) vorgesehenen Durchsuchermoduls (16) unter
23 Berücksichtigung von ermittelten Erfahrungswerten aus vorangegangenen
24 Suchanfragen gewählt werden.
25
 - 26 4. Verfahren nach einem der Ansprüche 1 bis 3, bei dem bei der Ermittlung der
27 Erfahrungswerte Korrelationen zwischen Einzelelementen der Suchanfragen
28 berücksichtigt werden.
29
 - 30 5. Verfahren nach einem der Ansprüche 1 bis 4, bei dem bei der Ermittlung der
31 Erfahrungswerte die relative Häufigkeit von Suchanfragen und/oder von Ein-
32 zelelementen der Suchanfragen berücksichtigt wird.

- 1
2 6. Verfahren nach Anspruch 5, bei dem denjenigen Datensätzen (4), die für ei-
3 ne vorgegebene Suchanfrage oder für eine vorgegebene Kombination aus
4 Einzelementen von Suchanfragen als relevant erkannt werden, proportio-
5 nal zur relativen Häufigkeit der Suchanfrage bzw. der Kombination aus Ein-
6 zelementen von Suchanfragen zusätzliche Systemressourcen für die Er-
7 stellung des zugeordneten Kurzdatensatzes (10) zugewiesen werden.
8
- 9 7. Verfahren zur Ermittlung von für eine vorgebbare Suchanfrage relevanten
10 Datensätzen (4) aus einer Datenbank, insbesondere aus dem World-Wide-
11 Web, bei dem in einem Speichermodul (6) hinterlegte, für die Datensätze (4)
12 charakteristische, nach dem Verfahren nach einem der Ansprüche 1 bis 6 er-
13 stellte Kurzdatensätze (10) auf ihre Relevanz für die jeweilige Suchanfrage
14 hin durchsucht werden.
15
- 16 8. Suchsystem (1) zur Ermittlung von für eine vorgebbare Suchanfrage relevan-
17 ten Datensätzen (4) aus einer Datenbank, insbesondere aus dem World-
18 Wide-Web, insbesondere zur Durchführung des Verfahrens nach Anspruch
19 7, bei dem in einem Speichermodul (6) für die Datensätze charakteristische
20 Kurzdatensätze (10) hinterlegt sind, wobei die zur Erstellung eines Kurzda-
21 tensatzes (10) aus einem Datensatz (4) bereitgestellten Systemressourcen
22 unter Berücksichtigung von hinterlegten Erfahrungswerten aus vorangegan-
23 genen Suchanfragen gewählt sind.
- 25 9. Suchsystem (1) nach Anspruch 8, bei dem die Erfahrungswerte eine für die
26 Anzahl gleichartiger Suchanfragen in einem vorgebbaren Zeitintervall cha-
27 rakteristische Kennzahl umfassen.
28
- 29 10. Suchsystem (1) nach Anspruch 8 oder 9, bei dem als Systemressourcen die
30 Ressourcen eines zur Erstellung der für die Datensätze (4) jeweils charakte-
31 ristischen Kurzdatensätze (10) vorgesehenen Durchsuchermoduls (16) unter
32 Berücksichtigung von hinterlegten Erfahrungswerten aus vorangegangenen
33 Suchanfragen gewählt sind.



**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☒ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.